



(51) International Patent Classification:

G10L 21/0216 (2013.01) G10L 15/26 (2006.01)

G10L 25/30 (2013.01)

(21) International Application Number:

PCT/EP2017/060637

(22) International Filing Date:

04 May 2017 (04.05.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/332622 06 May 2016 (06.05.2016) US

(71) Applicant: ROBERT BOSCH GMBH [DE/DE]; Postfach 30 02 20, 70442 Stuttgart (DE).

(72) Inventor: HASAN AL BANNA, Taufiq; 2952 Magliocco Drive, Apt. 14, San Jose, California 95128 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,

DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: SPEECH ENHANCEMENT AND AUDIO EVENT DETECTION FOR AN ENVIRONMENT WITH NON-STATIONARY NOISE

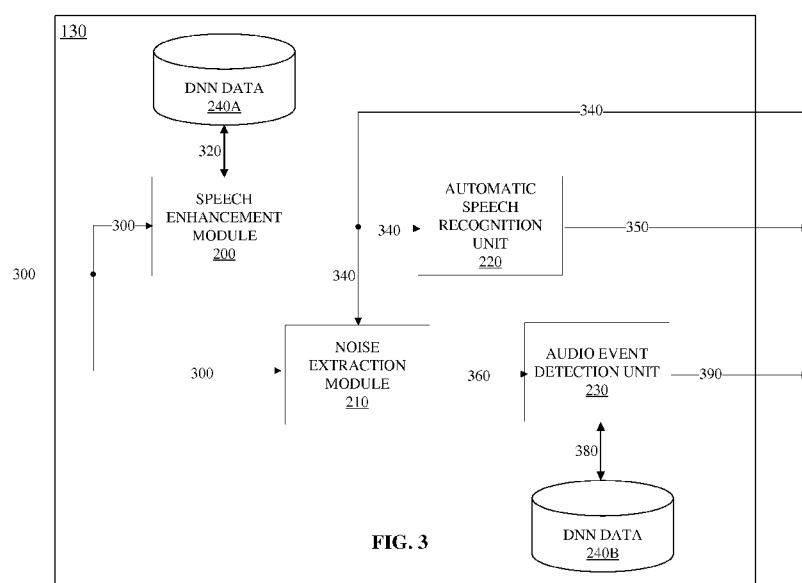


FIG. 3

(57) Abstract: A computer-implemented system and method provide an audio label for a noise signal. The computer-implemented system and method include receiving an audio input and obtaining the noise signal from the audio input. The computer-implemented system and method include extracting audio features of the noise signal. The computer-implemented system and method include determining and outputting an audio label for the extracted audio features of the noise signal based on machine learning data.

SPEECH ENHANCEMENT AND AUDIO EVENT DETECTION FOR AN ENVIRONMENT WITH NON-STATIONARY NOISE

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/332,622, which was filed on May 6, 2016, and which is hereby incorporated herein by reference in its entirety.

FIELD

[0002] This disclosure relates generally to acoustics, and more particularly to speech enhancement and audio event detection.

BACKGROUND

[0003] Unless otherwise indicated herein, the materials described in this section are not prior art to the claims in this application and are not admitted as prior art by inclusion in this section.

[0004] In an environment with background noise, a communication device with a microphone may transmit background noise components in addition to speech components. In these instances, the background noise components may interfere with the speech components. In turn, this interference may make it difficult for acoustic technology, such as automatic speech recognition technology (ASR) or other types of technologies, to provide effective results.

SUMMARY

[0005] The following is a summary of certain embodiments described in detail below. The described aspects are presented merely to provide the reader with a brief summary of these certain embodiments and the description of these aspects is not intended to limit the scope of this disclosure. Indeed, this disclosure may encompass a variety of aspects that may not be explicitly set forth below.

[0006] In an example embodiment, a computer-implemented method includes at least one computer processor, which is configured to obtain a noise signal from an audio input. The computer-implemented method includes extracting, via the at least one computer processor, audio features of the noise signal. The computer-implemented method includes determining and

outputting, via the at least one computer processor, an audio label for the audio features of the noise signal based on neural network data related to a first neural network model.

[0007] In an example embodiment, a computer system includes at least one computer processor configured to (a) obtain a noise signal from audio input, (b) extract audio features of the noise signal, and (c) determine and output an audio label for the audio features of the noise signal based on neural network data related to a first neural network model. The computer system includes at least one computer readable medium, which stores the neural network data of the first neural network.

[0008] These and other features, aspects, and advantages of the disclosure are further clarified by the following detailed description of certain exemplary embodiments in view of the accompanying drawings throughout which like characters represent like parts.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] FIG. 1A is an illustration of an exemplary speech enhancement and audio event detection system according to an example embodiment of the described disclosure.

[0010] FIG. 1B is an illustration of an exemplary speech enhancement and audio event detection system according to another example embodiment of the described disclosure.

[0011] FIG. 2 is an illustration of a non-limiting example of an implementation of the speech enhancement and audio event detection system according to an example embodiment of the described disclosure.

[0012] FIG. 3 is a schematic block diagram of the speech enhancement and audio event detection system according to an example embodiment of the described disclosure.

[0013] FIG. 4 is a schematic block diagram of the speech enhancement module, the noise determination module, and the audio event detection unit according to an example embodiment of the described disclosure.

[0014] FIG. 5 is a flow diagram of a training phase for speech enhancement according to an example embodiment of the described disclosure.

[0015] FIG. 6 is a flowchart of a method for extracting audio features according to an example embodiment of the described disclosure.

[0016] FIG. 7 is a flow diagram of a training phase for audio event detection according to an example embodiment of the described disclosure.

[0017] FIG. 8 is a flowchart of a method for extracting audio features according to an example embodiment of the described disclosure.

DETAILED DESCRIPTION

[0018] The embodiments described above, which have been shown and described by way of example, and many of their advantages will be understood by the foregoing description, and it will be apparent that various changes can be made in the form, construction and arrangement of the components without departing from the disclosed subject matter or without sacrificing one or more of its advantages. Indeed, the described forms of these embodiments are merely explanatory, these embodiments are susceptible to various modifications and alternative forms, and the following claims are intended to encompass and include such changes and not be limited to the particular forms disclosed, but rather to cover all modifications, equivalents, and alternatives falling within the spirit and scope of this disclosure.

[0019] FIG. 1A illustrates an example embodiment of a system 100A that includes at least one computing system 108, at least one speech enhancement and audio event detection system 130, and at least one server system 160 (or any suitable computer system 160). In addition to these components, the system 100A also includes at least one communication link 140, at least one communication network 146, and at least one communication link 150 such that the computing system 108 and the server system 160 are communicatively coupled to each other and such that the speech enhancement and audio event detection system 130 can receive audio input and provide the functions described herein.

[0020] In an example embodiment, the computing system 108 may be a personal computer or desktop computer, a laptop, a cellular or smart phone, a tablet, a personal digital assistant (PDA), a gaming console, an audio device, a video device, a diagnostic device, an entertainment device such as a television, a vehicle infotainment, a wearable device, a thin client system, a thick client

system, or the like. In some embodiments, the computing system 108 is a single client machine or a single group of client machines. In an example embodiment, as illustrated in FIG. 1A, the computing system 108 includes a computing device 120, e.g. a portable device such as a smart phone communicatively coupled to an audio device such as a headset 110. The headset 110 includes at least an acoustic sensor such as a microphone, a speaker, or any combination thereof. In another embodiment, the audio device 110 may communicatively couple directly to the communication network 146 or the server system 160. In this regard, the headsets 110 are not limited to those that are shown in FIGS. 1A and 1B, but may include any suitable wired, wireless, or any combination thereof that include any suitable number of acoustic sensors. As an example, the audio device may be a hearing aid, a listening device, goggles, or glasses with an integrated audio device, a helmet with an integrated audio device, or any suitable audio devices. In an example embodiment, the microphone includes directional, unidirectional, or omni-directional properties. In an example embodiment, the headset 110 has noise cancelling capability. In an example embodiment, the headset 110 is configured to cover ears of a user and attenuate external noise. In FIG. 1A, the headset 110 includes acoustic earmuffs to block out external noise and protect the ears of a user. In FIG. 1B, the headset includes noise cancelling capability.

[0021] In an example embodiment, the computing device 120 is in communication with the headset 110 via at least one communication link 140, which is wired, wireless, or any combination thereof. In an example embodiment, the headset 110 is in communication with at least its corresponding computing device 120, server system 160, communication network 146, or any combination thereof via suitable communication links 140 and/or 150, which can include Bluetooth® technology, WiFi technology, or any communication technology. The communication links 140 and 150 may be wired, wireless, or any combination thereof. The wireless communication links 140 and 150 may include cellular protocol, data packet protocol, radio frequency protocol, satellite band, infrared channel, or any other protocol able to transmit data among computing system 108, the server system 160, and the communication network 146. The wired communication link may include any wired line link.

[0022] In an example embodiment, the computing system 108 includes the appropriate network and/or communications technology such that the computing device 120 is enabled to communicate

with at least an audio device (e.g., the headset 110) and the speech enhancement and audio event detection system 130.

[0023] In an example embodiment, the speech enhancement and audio event detection system 130 is implemented via a computer system with at least one processor and at least one electronic storage device, as well as the appropriate network technology. For example, in FIG. 1A, the speech enhancement and audio event detection system 130 is implemented via the server system 160. In this regard, for example, the server system 160 includes at least one server computer and at least one electronic storage device, as well as the appropriate network technology to communicate with the computing device 120. In this case, since the speech enhancement and audio event detection system 130 is implemented via the server system 160, the system 100A includes at least one communication network 146.

[0024] In an example embodiment, the communication network 146 can include at least one computer network, at least one data network, at least one telecommunications network, any suitable network, or any combination thereof. Communication network 146 is not limited in its structure, configuration, and/or components as long as it is able to provide the necessary communications between the speech enhancement and audio event detection system 130 of the server system 160 and the computing system 108. The communication network 146 can comprise one or more sub-networks, and can be installed between any combination of the computing devices 108 and the server system 160. In some embodiments, the communication network 146 can be, for example, a local-area network (LAN), a metropolitan area network (MAN), a wide area network (WAN), a primary network comprising multiple sub-networks located between the computing devices 108 and the server system 160, a cloud network, or any combination thereof. Still further embodiments include the communication network 146 that can include any network types such as a point to point network, a broadcast network, a telecommunication network, a data communication network, a computer network, an ATM (Asynchronous Transfer Mode) network, a SONET (Synchronous Optical Network) network, a SDH (Synchronous Digital Hierarchy) network, a wireless network, a wireline network, and the like. Depending on the application, other networks may be used so that data exchanged between the computing system 108 and the server system 160. Network topology of the communication network 146 can differ within different embodiments which may include a bus network topology, a star network topology, a ring network

topology, a repeater-based network topology, or a tiered-star network topology. Additional embodiments may include a network of mobile telephone networks that use a protocol to communicate among mobile devices, where the protocol can be for example AMPS, TDMA, CDMA, GSM, GPRS, UMTS, LTE or any other protocol able to transmit data among mobile devices.

[0025] FIG. 1B illustrates an alternative example embodiment of a system 100B that includes at least one computing system 108, at least one speech enhancement and audio event detection system 130, and at least one communication link 140. As aforementioned, the computing system 108 includes at least one processor and at least one electronic storage medium, as well as the appropriate network/communications technology such that communications may occur between the appropriate hardware/software components of the computing system 108 and the speech enhancement and audio event detection system 130.

[0026] As discussed above, the speech enhancement and audio event detection system 130 can be implemented by various types of computer hardware and computer software, as long as the speech enhancement and audio event detection system 130 is configured to provide at least a clean speech signal 340, text 350, audio event data 390, or any combination thereof. In addition, the speech enhancement and audio event detection system 130 is configured to operate in any target environment in which machine learning techniques have been employed and the necessary machine learning data has been extracted. In this regard, the target environment itself can vary. For example, the target environment can be a manufacturing facility, a research lab, a construction site, or any environment in which a suitable amount of training data, such as audio samples corresponding to that target environment, is available.

[0027] FIG. 2 illustrates a non-limiting example of an implementation of the speech enhancement and audio event detection system 130 according to an example embodiment of the disclosure. In this particular example, the speech enhancement and audio event detection system 130 is configured to operate in an auto repair workshop environment. To be effective in this environment, the speech enhancement and audio event detection system 130 is configured to process the audio input 300 based on machine learning data (e.g., neural network model parameters) derived from audio samples, which are associated with an auto repair workshop environment.

[0028] In an example embodiment, the speech enhancement and audio event detection system 130 is configured to receive audio input 300. For example, in FIGS. 1A and 1B, the audio input 300 is transmitted to the speech enhancement and audio event detection system 130 via the headset 110. For instance, in this auto repair workshop example, a user has communicated his desire to receive auto-repair help via any suitable computer assistive technology, running on the computing device 120, by saying, "How do I replace the ignition system?" into the microphone of the headset 110.

[0029] Also, in this particular example, the audio input 300 includes some background noise from the auto-repair workshop. In this case, the background noise includes non-stationary noise such as a sound of a drill operating proximal to the user. That is, in this auto repair workshop example, the audio input 300, which is transmitted to the speech enhancement and audio event detection system 130, includes at least one speech component (e.g., "How do I replace the ignition system?") and at least one noise component (e.g., the sound of a drill).

[0030] In an example embodiment, the speech enhancement and audio event detection system 130 is configured to receive and process this audio input 300 in real-time. Also, in an example embodiment, the speech enhancement and audio event detection system 130 is configured to deliver clean speech 340 in real-time. For instance, in the auto repair workshop example, the user's voiced input of "How do I replace the ignition system?" is separated from the background noise (e.g., the sound of the drill) and delivered as clean speech 340 in real-time.

[0031] In an example embodiment, the speech enhancement and audio event detection system 130 is configured to provide at least the recognition and translation of the clean speech 340 into text 350. In the auto repair workshop example, the user's voiced input of "How do I replace the ignition system" is converted into the following text 350: "How do I replace the ignition system?" for display on computing system 108, any suitable device on the communication network 146, the server system 160, or any combination thereof.

[0032] Furthermore, in an example embodiment, the speech enhancement and audio event detection system 130 is configured to provide at least audio event data 390 relating to at least one identified noise component of the audio input 300. For example, in the auto repair workshop

example, the speech enhancement and audio event detection system 130 is configured to detect the sound of the drill from the audio input 300 and provide an identification label of "DRILL."

[0033] Audio event data 390, such as the identification label of "DRILL," provides contextual information for the clean speech 340 ("How do I replace the ignition system") and/or the text 350. This contextual information provides a number of advantages. In this case, for example, upon being notified that a "DRILL" is being detected as an audio event in real-time, the computer-assistive technology is provided with contextual information, indicating that the drill is most likely in use even though the user has asked about replacing the ignition system. In this case, the computer-assistive technology is able to provide a message, such as "Please complete the drill operation started in the previous repair step before moving on to ignition replacement." In this regard, due to the contextual information provided by the audio event data 390, computer assistive technology, running on device 120, is able to provide safety notifications to the user. In addition, in such a case, the computer assistive technology is able to wait until the drill is no longer detected before providing the user with the repair steps on how to replace the ignition system. In this regard, by being provided with this audio event data 390, the computer assistive technology is able to ensure that the appropriate tools and steps are taken to perform the repair operation.

[0034] As another non-limiting example, the user could be in the middle of a repair and ask "what is the next step" into the headset 110. In this scenario, in addition to user's question, the speech enhancement and audio event detection system 130 detects the sound of a grinder tool as background noise and provides the audio event data 390 of "GRINDER." Meanwhile, the clean speech signal 340 is input to the ASR unit 220 and converted into text 350. In this example, the text 350 is processed by the Natural Language Understanding (NLU) engine, thereby enabling the computer assistive technology to understand that the user wants to know the next step of the repair operation. Given the contextual information that the grinder is still running, the computer assistive technology is configured to provide the following prompt: "Please stop the GRINDER and use an IMPACT WRENCH for the next repair step." In this regard, by receiving at least the text 350 together with the audio event data 390 from the speech enhancement and audio event detection system 130, the computer assistive technology is able to respond immediately to the user with the appropriate notifications.

[0035] FIGS. 3-4 are block diagrams of the speech enhancement and audio event detection system 130 according to an example embodiment. More specifically, FIG. 3 illustrates an overview of the speech enhancement and audio event detection system 130, whereas FIG. 4 illustrates certain features of the speech enhancement and audio event detection system 130 in greater detail. As shown, the speech enhancement and audio event detection system 130 is configured to receive audio input 300, which can be any type of audio signal. The speech enhancement and audio event detection system 130 is configured to process the audio input 300 and provide clean speech 340, text 350, audio event data 390, or any combination thereof. In an example embodiment, the speech enhancement and audio event detection system 130 includes a speech enhancement module 200, a noise extraction module 210, an automatic speech recognition (ASR) unit 220, an audio event detection (AED) unit 230, deep neural network (DNN) data 320, and DNN data 380. In an alternative embodiment, the ASR unit 220 is omitted from being a component within the speech enhancement and audio event detection system 130, but optionally connected to the speech enhancement and audio event detection system 130.

[0036] In an example embodiment, the speech enhancement module 200 is configured to receive audio input 300 and access DNN data 320 from at least one electronic storage device 240A. Based on the audio input 300 and the DNN data 320, the speech enhancement module 200 is configured to generate a clean speech estimate in real-time. The clean speech estimate is an audio signal in which noise of the audio input 300 is removed or reduced. In an example embodiment, the clean speech estimate is a clean speech signal 340.

[0037] In an example embodiment, to provide the clean speech signal 340, the speech enhancement module 200 includes at least an audio feature extractor 202, a DNN decoding unit 204, and a waveform reconstruction unit 206. In an example embodiment, the audio feature extractor 202 is configured to receive the audio input 300 in real-time. Upon receiving the audio input 300, the audio feature extractor 202 is configured to extract certain audio features 310 from the audio input 300. In this regard, for example, the audio features 310 can include log power spectrum (LPS) parameters, Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) coefficients, other suitable acoustic/audio features, signal data, or any combination thereof. More specifically, the audio features, which are extracted by the audio feature extractor 202, are the same audio features that

are extracted during a training phase (FIG. 5) to obtain the DNN data 320. For example, if LPS data was extracted during the training phase to obtain the DNN data 320, then the audio feature extractor 202 extracts LPS data from the audio input 300. Once extracted, the audio features 310 are transmitted to the DNN decoding unit 204.

[0038] In an example embodiment, the DNN decoding unit 204 is configured to receive the audio features 310 from the audio feature extractor 202. Upon receiving the audio features 310, the DNN decoding unit 204 is configured to decode the audio features 310 using DNN data 320, which is accessible from at least one electronic storage device 240A. More specifically, in an example embodiment, the DNN data 320 includes relationship information between various clean and noisy audio features. Based on this DNN data 320, the DNN decoding unit 204 is configured to obtain the clean audio features 330 from the audio features 310, which were extracted from the audio input 300. Next, the clean audio features 330 are transmitted from the DNN decoding unit 204 to the waveform reconstruction unit 206.

[0039] In an example embodiment, the waveform reconstruction unit 206 is configured to perform waveform reconstruction based on the clean audio features 330 in real-time. More specifically, the waveform reconstruction unit 204 is configured to receive the clean audio features 330 from the DNN decoding unit 204 and generate a clean speech signal 340 (or a clean speech estimate 340) from the clean audio features 330. Afterwards, in an example embodiment, the clean speech signal 340 is transmitted from the waveform reconstruction unit 206 to the noise extraction module 210. Additionally or alternatively, in an example embodiment, the clean speech signal 340 is transmitted from the waveform reconstruction unit 206 to the ASR unit 230. Additionally or alternatively, in an example embodiment, the clean speech signal 340 is provided as output from the waveform reconstruction unit 206.

[0040] In an example embodiment, the ASR unit 220 includes speech recognition technology to provide speech recognition and translation services of the clean speech signal 340. For example, in the example shown in FIG. 3, the speech enhancement and audio event detection system 130 includes an ASR unit 220, which is configured to translate the clean speech signal 340 into text 350. The ASR unit 220 is configured to implement an algorithm that automatically converts speech signals into text 350. For example, the ASR unit 220 is configured to include and/or

execute any suitable ASR software or commercially available software, such as that provided by Google, Nuance, AT&T, or any other entity. As another example, the ASR unit 220 is configured to include and/or execute any open source software toolkits, which are available to build ASR algorithms, such as that provided by HTK, Kaldi, or any other entity.

[0041] In an example embodiment, the noise extraction module 210 includes a noise extractor 212, which is configured to extract a noise signal 360 (or a noise estimate 360) from the audio input 300. In addition, the noise extraction module 210 is configured to provide the extracted noise signal 360 to the AED unit 230. To do so, the noise extractor 212 is configured to receive both the audio input 300 and the clean speech signal 340. Upon receiving these inputs, the noise extractor 212 is configured to extract the non-speech portion, such as the noise signal 360, from the audio input 300 by removing the clean speech signal 340 from the audio input 300.

[0042] More specifically, in an example embodiment, the noise extractor 212 is configured to find the value of 'b' such that $E_t[(y(t) - b s'(t))^2]$ is minimized, where $y(t)$ represents the audio input 300 taken at time "t", $s'(t)$ represents the clean speech signal 340 taken at time "t," and E_t is the expectation operator over time "t." In this regard, the noise extractor 212 is configured to obtain $n'(t)$, which represents the noise signal 360, for example, via Equation 1.

[0043]
$$n'(t) = y(t) - b s'(t) \quad \text{[Equation 1]}$$

[0044] In an example embodiment, the noise extractor 212 is configured to output the noise signal 360 (or the noise estimate 360) in real-time. In some instances, the noise estimate 360 includes background noise, such as a stationary noise, a non-stationary noise, or both stationary and non-stationary noises. Afterwards, the noise signal 360 is transmitted to the AED unit 230.

[0045] In an example embodiment, the AED unit 230 is configured to receive the noise estimate 360 from the noise extraction module 210. In addition, the AED unit 230 is configured to access DNN data 380 from at least one electronic storage device 240B upon receiving the noise estimate 360. Based on these inputs, AED unit 230 is configured to detect and recognize audio events. In addition, the AED unit 230 is configured to provide audio event data 390 based on this detection and recognition in real-time. For example, referring to FIG. 2, when the target environment is an auto repair workshop, the AED unit 230 is configured to provide identification labels that identify

a tool, an auto-repair sound, or any noise, which is detected from the noise signal 360 (or the noise estimate 360) of the audio input 300. To do so, in an example embodiment, as shown in FIG. 4, the AED unit 230 includes an audio feature extractor 232 and a DNN decoding unit 234.

[0046] In an example embodiment, the audio feature extractor 232 is configured to receive the noise signal 360. Upon receiving the noise signal 360, the audio feature extractor 232 is configured to extract certain audio features 370 from the noise signal 360. In this regard, for example, the audio features 370 can include log power spectrum (LPS) parameters, Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) coefficients, other acoustic/audio features, suitable signal data, or any combination thereof. More specifically, the audio features, which are extracted by the audio feature extractor 232, are the same audio features that are extracted during a training phase (FIG. 7) to obtain the DNN data 380. For example, if MFCC was extracted during the training phase to obtain DNN data 380, then the audio feature extractor 232 extracts MFCC from the noise signal 360. Once extracted, the audio features 370 are transmitted to the DNN decoding unit 234.

[0047] In an example embodiment, the DNN decoding unit 234 is configured to receive the audio features 370 from the audio feature extractor 232. Upon receiving the audio features 370, the DNN decoding unit 234 is configured to decode the audio features 370 of the noise signal 360 by using DNN data 380, which is accessible from at least one electronic storage device 240B. Based on the decoding, the DNN decoding unit 234 is configured to provide audio event data 390.

[0048] In an example embodiment, when performing the decoding operation, the DNN decoding unit 234 is configured to select the audio event data 390, which most likely corresponds to the extracted audio features. In this regard, for example, the DNN decoding unit 234 is configured to provide different audio labels and likelihood scores. For instance, with regard to the auto repair workshop scenario discussed above, a non-limiting example of some of the audio labels and likelihood scores, which are provided by the DNN decoding unit 234, are as follows: "DRILL": 1.5, "WRENCH": -0.5, "HAMMER": -1.2, and etc. As demonstrated by this example, the values of the likelihood scores can range anywhere from negative values to positive values. In addition, the likelihood scores are obtained for all audio event data 390 (e.g., all of the audio labels of FIG. 7). Accordingly, in this example embodiment, the audio label with the highest likelihood score or

the maximum likelihood score is determined to most likely be the corresponding audio event for the audio features 370. For instance, in the above example, the audio label of “DRILL” is determined to have the highest likelihood score of 1.5 and is therefore selected by the DNN decoding unit 234 to be the audio event data 390. That is, in this scenario, the DNN decoding unit 234 outputs “DRILL” as the audio event data 390.

[0049] In an example embodiment, the likelihood score is calculated from DNN output values in the following manner. First, the DNN output values, denoted by y_i , where $i = 1, 2 \dots N$ (corresponding to the “N” number of audio labels), are passed to a softmax function to calculate corresponding z_i values, for example, via Equation 2.

$$z_i = \frac{e^{y_i}}{\sum_k^N e^{y_k}} \quad \text{[Equation 2]}$$

[0050]

[0051] In an example embodiment, the softmax function converts the DNN output, i.e., y_i values, to a set of z_i values so that each of them are between 0 to 1, and so they add up to 1, i.e., $z_1 + z_2 \dots z_N = 1$. Next, for each of these z_i values, the likelihood score L_i is calculated, for example, via Equation 3.

$$L_i = \log \left(\frac{z_i}{\sum_{j=1, j \neq i}^N z_j} \right) \quad \text{[Equation 3]}$$

[0052]

[0053] As an example, suppose the DNN output includes at least the following three values: y_1 , y_2 and y_3 . These values correspond to the audio labels of “DRILL”, “WRENCH” and “HAMMER”, respectively. If $y_1 = 0.9$, $y_2 = 0.2$, and $y_3 = 0.1$, then after the softmax operation, the DNN decoding unit 234 is configured to determine that $z_1 = 0.52$, $z_2 = 0.25$, and $z_3 = 0.23$. In this case, the DNN decoding unit 234 determines that the likelihood scores are $L_1 = 0.03$, $L_2 = -0.47$, $L_3 = -0.52$.

[0054] In this example embodiment, the DNN decoding unit 234 does not provide a “likelihood score” for a noise signal (or audio event), which it does not observe during the training phase (FIG. 7). In this regard, for example, if all of the likelihood scores are below a certain value or a predetermined threshold (e.g., zero), then the AED unit 230 is configured to determine that none of the audio events (e.g., audio labels) are likely to correspond to the noise signal 360. In this

scenario, the AED unit 230 considers the audio event data 390 to be NULL. Also, in an example embodiment, when a noise occurs which does not correspond to any audio events that were used during the training phase (FIG. 7), the DNN decoding unit 234 is configured to provide a low likelihood score for all of the existing labels below zero, as described earlier. In this case, the DNN decoding unit 234 decides that none of the audio events are "likely enough" and sets the output of the audio event data 390 to be NULL.

[0055] Additionally or alternatively, in an example embodiment, the DNN decoding unit 234 is configured to consider high likelihood scores from two different audio labels (e.g., a likelihood score of 1.5 for "DRILL" and a likelihood score of 1.4 for "HAMMER") and decide that both of the audio events have occurred. In some cases, although a second audio label is detected from the audio input 300 and determined to be a likely audio event, this second audio label (e.g., "HAMMER"), which is detected, may not impact the current repair scenario and may not change the systems prompts provided by computer assistive technology of the computing device 120. The DNN decoding unit 234 is configured to perform DNN decoding and provide likelihood scores for each of the audio events (e.g., audio labels) in real-time.

[0056] As discussed above, DNN data 320 is accessed during the process of providing the clean speech signal 340, whereas DNN data 380 is accessed during the process of providing audio event data 390. Also, as discussed, DNN data 320 and DNN data 380 are stored in at least one electronic storage device 240. In an example embodiment, the reference labels 240A and 240B may refer to different portions of the same electronic storage device. In an alternative example embodiment, each of the reference labels 240A and 240B refers to at least one electronic storage device such that the reference labels 240A and 240B comprise a plurality of electronic storage devices. Furthermore, irrespective of their electronic configurations and storage arrangements, DNN data 320 and DNN data 380 are provided to at least one electronic storage device after being obtained during training phases (FIGS. 5 and 7).

[0057] In an example embodiment, the training phases occur when the computer system, which is implementing the speech enhancement and audio event detection system 130, is offline. In this regard, for example, the training phases may occur when the server system 160 or the computing device 120 is offline. A detailed description of the training phases is discussed below.

[0058] FIG. 5 is a flow diagram of a training phase for speech enhancement according to an example embodiment. The training phase includes extracting audio features from audio samples. In order to be effective, the DNN training module is trained with a large number of audio samples associated with a target environment. As shown, the DNN training module may be trained with 'N' number of audio samples, where 'N' represents a suitable number of samples. For example, in an example embodiment, 'N' represents an integer number that is equal to or greater than 10,000 audio files (with a total duration of 500 hours). In general, the speech enhancement module 200 is configured to perform better when utilizing DNN data 320, which is obtained by training a DNN model with a greater number of samples.

[0059] In an example embodiment, each audio sample includes a clean speech sample and a noisy speech sample. Each noisy speech sample is obtained by adding a noise sample to a clean speech sample. More specifically, the noisy speech sample may be represented by Equation 4.

[0060]
$$y(t) = s(t) + n(t) \quad \text{[Equation 4]}$$

[0061] In this equation, $y(t)$ represents the noisy speech sample taken at time t , $s(t)$ represents the clean speech sample taken at time t , and $n(t)$ represents the noise sample taken at time t . As indicated, with respect to time t , the noisy speech sample equates to a sum of a clean speech sample and a noise sample. In this regard, various combinations of clean speech samples, noise samples, and noisy speech samples are generated for use as audio samples.

[0062] For instance, in the auto repair workshop example, the audio samples may include noisy speech samples in which the noise samples are typical sounds of an auto repair workshop. In this regard, for example, the noise samples include audio recordings of non-stationary noises, stationary noises, or any combination thereof. As examples, the noise samples may include audio recordings of a drill, a cutter, a welding system, other auto repair workshop sounds, or any combination thereof. In this regard, the DNN training module is configured to employ machine learning techniques to learn a variety of noises of the auto-repair workshop and to generate DNN data 320 based thereupon.

[0063] In an example embodiment, the DNN training module includes a DNN model with multiple layers. For example, the DNN model includes an input layer, hidden layers, and an output layer.

As shown in FIG. 5, the input layer includes a plurality of input nodes (e.g., X_1 to X_k , where 'k' represents the number of nodes) and the output layer includes a plurality of output nodes (e.g., Y_1 to Y_m , where 'm' represents the number of nodes). In addition, each of the hidden layers includes a plurality of nodes (e.g., H_1 to H_n , where 'n' represents the number of nodes). More specifically, in an example embodiment, the DNN model includes an input feature dimension of 350 in which the input feature dimension is equal to the number of input nodes. Alternatively, in a case in which a sequence of feature vectors are passed to the DNN, the feature vector dimension and the number of DNN input nodes can be different. Meanwhile, in this example, the DNN model includes 3 hidden layers in which each hidden layer includes 2500 nodes. Also, in this example, the DNN model includes 350 output nodes.

[0064] In an alternative example embodiment, the DNN model includes an input layer with multiple audio features as input. For example, if an audio feature, which is represented by a vector, contains "D" number of parameters and if "N" number of consecutive audio features (coming from the audio samples in real-time) are input into the DNN model at the same time, then the input layer includes $D \times N$ nodes, i.e. the input layer includes a number of nodes that is equal to D multiplied by N.

[0065] In an example embodiment, each node of the DNN model takes a scalar number as input in the input layer. Each scalar number is transmitted to corresponding nodes in a hidden layer with different weight parameters "W." For each node in each layer, the weight parameters "W" are unique. Each node in a hidden layer accepts the corresponding output values from the previous layer and adds them to form one scalar parameter. This parameter is then processed through a sigmoid function $S(t)$, as set forth, for example, in Equation 5.

[0066]
$$S(t) = \frac{1}{1+e^{-t}} \quad \text{[Equation 5]}$$

[0067] The corresponding result from the sigmoid function is then passed to the next layer of nodes. The process of DNN training generally involves learning the weight parameters "W" such that when a noisy audio feature is provided as input, then the output closely approximates a clean version of that audio feature.

[0068] In an example embodiment, the DNN training module includes a stack of restricted Boltzmann machines (RBMs) to initialize the DNN model. The structure of the RBM layers are the same or substantially similar to the DNN layers except that it follows an unsupervised training mechanism. The RBMs provide pre-training for the DNN training. In this regard, the RBMs are a mathematical tool, which is implemented in software and used to train the DNN model.

[0069] In an example embodiment, the DNN training module includes a DNN model with an input layer associated with the noisy speech samples and an output layer associated with the clean speech samples. In this regard, for instance, the input layer receives the noisy audio features (X_{noisy}) and the desired output is the corresponding clean audio features (Y_{output}). To achieve this, the RBM pre-training, as discussed above, is performed. Next, for each noisy audio feature frame that is input, the output is calculated as Y_{output} . This Y_{output} is then compared with the corresponding clean feature frame and an error is generated, as indicated, for example, by Equation 6.

[0070]
$$\text{Error} = E[(X_{\text{clean}} - Y_{\text{output}})^2] \quad [\text{Equation 6}]$$

[0071] Equation 6 is a mean squared error function between the audio features of the output layer (Y_{output}) and desired clean features (X_{clean}). As indicated above, the error, calculated for this equation, is lower when the clean audio features are more similar to the audio features of the output. This error is then used with the back-propagation method, which finds the responsibility of each of the "W" parameters in the DNN model for this error and modifies them so that the error is reduced. In an example embodiment, these operations are iterated many times (e.g., at least 100 times).

[0072] In an example embodiment, DNN training module is configured to implement various machine learning techniques (e.g., regression based machine learning techniques), functions, operations, and equations such that its DNN model learns how to obtain the audio features of clean speech given the audio features of the noisy speech. The DNN training module is configured to store its DNN model parameters as DNN data 320 in at least one electronic storage device 240A. In an example embodiment, the DNN data 320 includes at least the weight parameters "W" (FIG. 5), which are obtained through the iterative training process of the DNN model, as discussed above.

[0073] FIG. 6 is a flowchart of a method for extracting audio features according to an example embodiment. This extraction process is configured to be implemented as (i) the audio feature extraction process performed by the audio feature extractor 202, (ii) the audio feature extraction process performed by the audio feature extractor 232, (iii) the audio feature extraction process performed in FIG. 5, (iv) the audio feature extraction process performed in FIG. 7, or (v) any combination thereof. To illustrate, in an example embodiment, this extraction process is performed at least during the training phase in which audio features are extracted in FIG. 5, as well as during normal operation when the audio features are extracted by the audio feature extractor 202.

[0074] As shown, the method includes extracting audio features from an input such as an audio signal at 600. For example, when this method is applied during the training phase, the input may refer to each of the audio samples, as shown in FIG. 5. In this regard, the method of extracting audio features is applied to the clean speech samples and the noisy speech samples. As another example, when this method is applied by the audio feature extractor 202, as shown in FIG. 4, the input may refer to audio input 300.

[0075] Once the input is received, the method includes framing the input as indicated at 602, to provide audio frames. The audio frames may be a suitable length. For example, the audio frames may be 25 milliseconds in duration. At 604, the method includes performing a discrete Fourier transform (DFT) on the audio frames. At 606, the method includes calculating LPS data based on the DFT. More specifically, the LPS data includes LPS coefficients, which are calculated from DFT coefficients by taking the magnitude squared and by taking the logarithm of the values. Once the calculations have been performed, the method includes providing the LPS data as output, as indicated at 608.

[0076] FIG. 7 is a flow diagram of a training phase for audio event detection according to an example embodiment. The training phase includes extracting audio features from audio samples. As shown, the DNN training module may be trained with 'N' number of audio samples, where 'N' represents a suitable number of samples. For example, in an example embodiment, 'N' represents an integer number that is equal to or greater than 10,000 audio files (with a total duration of 500 hours). In this regard, for example, the N audio samples are the same audio samples, which were

discussed with respect to FIG. 5. In general, the AED unit 230 is configured to perform better when utilizing DNN data 380, which is obtained by training a DNN model with a greater number of samples.

[0077] Once extracted, the audio features are transmitted to the DNN training module. Additionally, as shown, the DNN training module is configured to receive audio event data 390, such as audio labels, which relate to the audio samples and possible audio events of the target environment. In an example embodiment, the audio labels include identification labels, audio descriptions, other relevant information, or a combination thereof.

[0078] For instance, in the auto repair workshop example, the audio labels may include the names of the auto repair tools, the auto repair sounds, sources of noise, or any suitable audio information. In this regard, for instance, one of the audio samples may include a noisy speech sample, which includes speech together with a hammering sound. In this case, the audio labels include an audio label of "HAMMER" to correspond to the hammering sound.

[0079] In an example embodiment, the DNN training module includes a DNN model with multiple layers. For example, the DNN model includes an input layer, hidden layers, and an output layer. As shown in FIG. 5, the input layer includes a plurality of input nodes (e.g., X_1 to X_K , where 'K' represents the number of nodes) and the output layer includes a plurality of output nodes (e.g., Y_1 to Y_N , where 'N' represents the number of nodes). In addition, each of the hidden layers includes a plurality of nodes (e.g., H_1 to H_M , where 'M' represents the number of nodes). More specifically, in an example embodiment, the DNN model for speech includes an input feature dimension for MFCC of 60 in which the number of input nodes is equal to the input feature dimension that is fed into the DNN model. In this case, the DNN model includes 2 hidden layers in which each hidden layer includes 100 nodes. In addition, in this example, the DNN model includes a number of nodes, which corresponds to the number of audio events, which are to be detected for the target environment. As an example, a target environment with 50 audio events (or audio labels) includes 50 nodes in its output layer.

[0080] As shown in FIG. 7 and discussed above, the DNN training module is configured to implement supervised machine learning techniques, such as regression based DNN, so that its DNN model learns how to identify different audio events and provide the appropriate audio labels

based on the extracted audio features. The DNN training module is configured to store its DNN model parameters as DNN data 380 in at least one electronic storage device 240B. In an example embodiment, the DNN data 380 includes at least the weight parameters “W” (FIG. 7), which are obtained through the iterative training process of the DNN model.

[0081] In an example embodiment, each weight parameter W is a value, which assigns the multiplying factor between two nodes. For example, in FIG. 7, the connection arrow between node X_1 and node H_{11} has a weight of W_{111} . In this regard, for example, the output from node X_1 will be multiplied by W_{111} before entering node H_{11} . In the example shown in FIG. 7, the set of W parameters includes three indices: j = the layer number, K = the node number from the source node, and M = the node number at the destination layer. Thus, for example, the output from node X_5 that goes into node H_{12} will have a weight parameter W_{152} , since, here: $i = 1$ (since it’s the input, or first layer), $K = 5$, and $M = 2$.

[0082] FIG. 8 is a flowchart of a method for extracting audio features according to an example embodiment. This extraction process is configured to be implemented as (i) the audio feature extraction process performed by the audio feature extractor 202, (ii) the audio feature extraction process performed by the audio feature extractor 232, (iii) the audio feature extraction process performed in FIG. 5, (iv) the audio feature extraction process performed in FIG. 7, or any combination thereof. For example, in an example embodiment, this method may be implemented at least during the training phase in which audio features are extracted in FIG. 7, as well as during normal operation when audio features are extracted by the audio feature extractor 232.

[0083] As shown, the method includes receiving input such as an audio signal at 700. For example, when this method is applied during the training phase, the input may refer to the audio samples, as shown in FIG. 7. As another example, when this method is applied by the audio feature extractor 232, as shown in FIG. 4, the input may refer to noise signal 360 (or noise estimate 360).

[0084] Once the input is received, the method includes framing the input, as indicated at 802, to provide audio frames. The audio frames may be any suitable length. In one example, the audio frames may be 25 milliseconds in duration. At 804, the method includes performing DFT on the audio frames. At 806, the method includes calculating LPS data based on the DFT. More specifically, the LPS data includes LPS coefficients, which are calculated from DFT coefficients

by taking the magnitude squared and by taking the logarithm of the values. Once the calculations have been performed, the method includes performing a filter bank analysis on the LPS data, as indicated at 808. Next, at 810, the method includes performing a discrete cosine transform (DCT) in accordance with results of the filter-bank analysis. Once the DCT has been performed, the method includes providing MFCC data as output, as indicated at 812.

[0085] As discussed above, the speech enhancement and audio event detection system 130 provides a number of advantages. For example, the speech enhancement and audio event detection system 130 is configured to provide clean speech 340 and/or text 350 from the audio input 300 even if background noise is present. In addition, by providing the speech enhancement and audio event detection system 130 with the ability to remove the clean speech from the audio input 300, the AED unit 230 is able to perform better by directly receiving a noise signal 360 instead of audio input 300.

[0086] In addition, the speech enhancement and audio event detection system 130 is configured to provide contextual information, such as audio event data 390. In effect, this contextual information benefits the user and any other system that is in communication with the speech enhancement and audio event detection system 130 by enabling computer assistive technology to provide intelligent, immediate, and customized services to the user based on the user's utterances and the context thereof. For instance, in the aforementioned auto repair workshop example, if a user is using the wrong tool while relying on computer assistive technology to perform a repair, then the computer assistive technology can immediately warn the user that the wrong tool is being used by recognizing the audio event (e.g., the sound of the wrong tool) from the background noise and relaying that information to the computer assistive technology and the user. As another example, upon detecting that the user is using a tool for a period that is longer than that which is recommended (or that an audio event is detected at an inappropriate time), the computer assistive technology can provide an appropriate message to the user. Also, as another example, if the audio event, which is detected, is determined to be too loud, then the computer assistive technology can prompt the user, for instance, to wear earplugs. As demonstrated by these illustrative examples and as discussed above, the speech enhancement and audio event detection system 130 provides a number of advantageous features.

[0087] Furthermore, as aforementioned, the AED unit 230 itself advantageously provides contextual information in the form of audio event data 390. Moreover, the advantages of the AED unit 230 can be obtained without being a part of the speech enhancement and audio event detection system 130. For example, in an alternative embodiment, the AED unit 230 is a standalone unit with access to the DNN data 380 stored in at one electronic storage device 240B. In this alternative embodiment, the AED unit 230 is configured to provide audio event data 390 by decoding audio features of its input based on the DNN data 380. That is, in this alternative embodiment, the AED unit 230 is configured to provide audio event data 390 without the additional processing of the speech enhancement module 200 and the noise extraction module 210. This is particularly advantageous in instances in which speech enhancement and speech recognition are not necessary, but fast processing of noise data and fast obtainment of audio event data are desirable. In addition, the AED unit 230, as a standalone unit, requires less hardware and/or software than that of the example embodiment shown in FIGS. 3-4.

[0088] That is, the above description is intended to be illustrative, and not restrictive, and is provided in the context of a particular application and its requirements. Those skilled in the art can appreciate from the foregoing description that the present invention may be implemented in a variety of forms, and that the various embodiments may be implemented alone or in combination. Therefore, while the embodiments of the present invention have been described in connection with particular examples thereof, the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the described embodiments, and the true scope of the embodiments and/or methods of the present invention are not be limited to the embodiments shown and described, since various modifications will become apparent to the skilled practitioner upon a study of the drawings, specification, and following claims. For example, components and functionality may be separated or combined differently than in the manner of the various described embodiments, and may be described using different terminology. These and other variations, modifications, additions, and improvements may fall within the scope of the disclosure as defined in the claims that follow.

What is claimed is:

1. A computer-implemented method comprising:
 - obtaining, by at least one computer processor, a noise signal from audio input;
 - extracting, by the at least one computer processor, audio features of the noise signal;
 - determining and outputting, by the at least one computer processor, an audio label for the audio features of the noise signal based on neural network data related to a first neural network model.
2. The computer-implemented method of claim 1, wherein the audio label identifies a noise source or an audio event that is associated with the noise signal.
3. The computer-implemented method of claim 1, wherein the neural network data is obtained by training the first neural network model with (i) audio samples associated with a target environment, and (ii) audio labels associated with the audio samples.
4. The computer-implemented method of claim 1, further comprising:
 - obtaining, by the at least one computer processor, a clean speech signal from the audio input,
 - wherein:
 - the noise signal is obtained by removing the clean speech signal from the audio input.
5. The computer-implemented method of claim 4, further comprising:
 - extracting, by the at least one computer processor, audio features of the audio input; and
 - determining and outputting, by the at least one computer processor, the clean speech signal for the audio features of the audio input based on neural network data related to a second neural network model.

6. The computer-implemented method of claim 5, wherein the neural network data of the second neural network model is obtained by training the second neural network model with audio samples associated with a target environment.
7. The computer-implemented method of claim 4, further comprising:
recognizing and translating, by the at least one computer processor, the clean speech signal into text.
8. A computer system comprising:
at least one computer processor configured to (a) obtain a noise signal from audio input, (b) extract audio features of the noise signal, and (c) determine and output an audio label for the audio features of the noise signal based on neural network data related to a first neural network model; and
at least one computer readable medium that stores the neural network data of the first neural network.
9. The computer system of claim 8, wherein the audio label identifies a noise source or an audio event that is associated with the noise signal.
10. The computer system of claim 8, wherein the neural network data of the first neural network model is obtained by training the first neural network model with (i) audio samples associated with a target environment, and (ii) audio labels associated with the audio samples.
11. The computer system of claim 8, wherein:
the at least one computer processor is configured to obtain a clean speech signal from the audio input, and
the at least one computer processor is configured to obtain the noise signal by removing the clean speech signal from the audio input.
12. The computer system of claim 11, wherein:

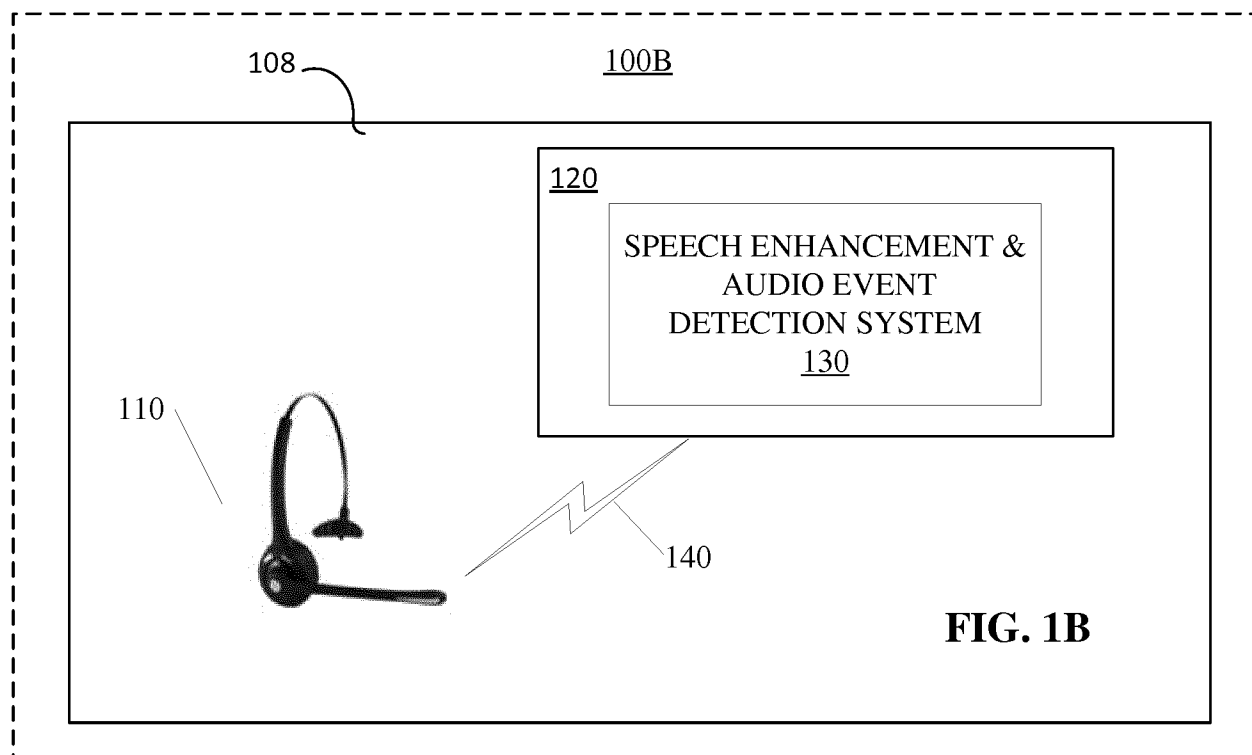
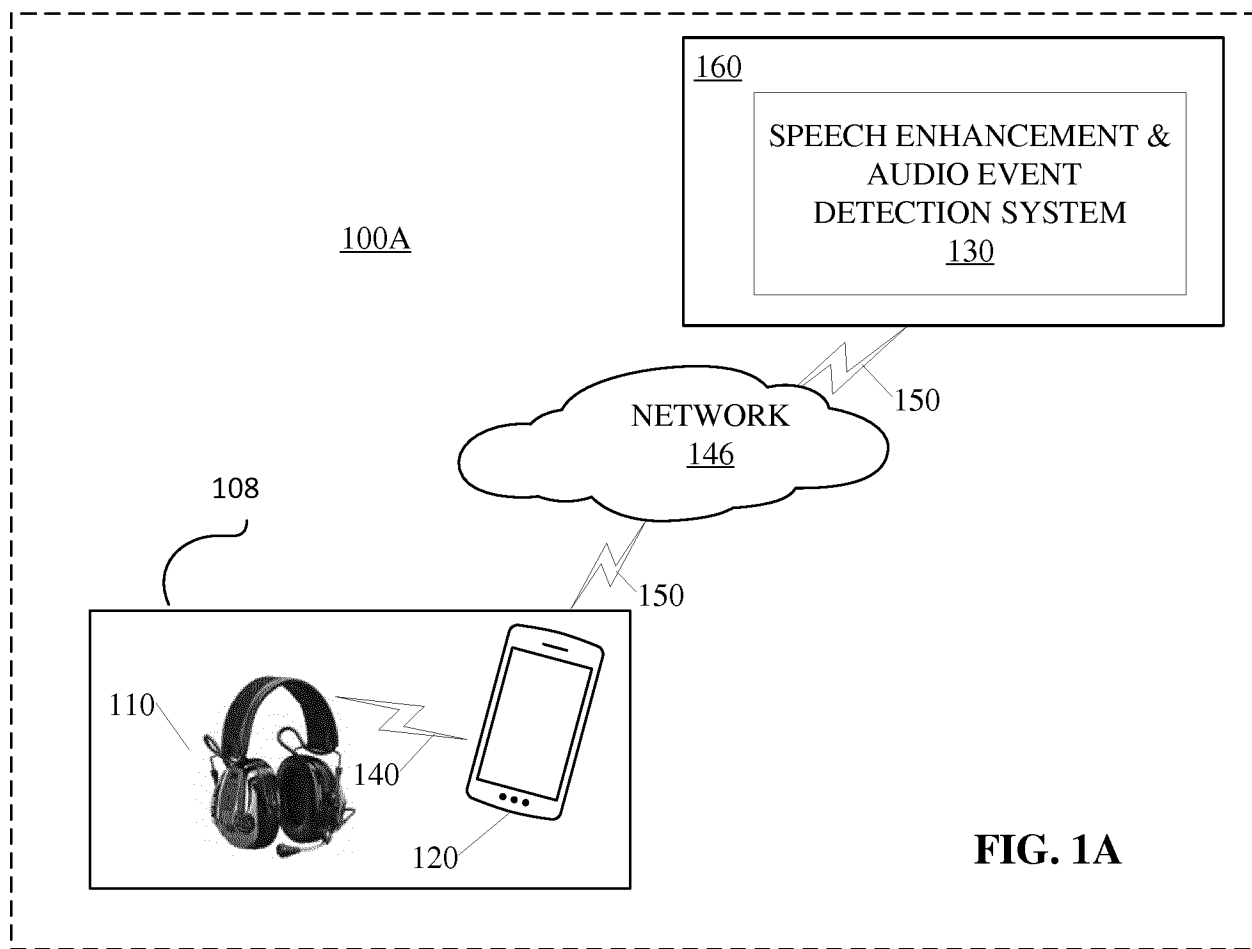
the at least one computer processor is configured to extract audio features of the audio input, and

the at least one computer processor is configured to determine and output the clean speech signal for the audio features of the audio input based on the neural network data of a second neural network model.

13. The computer system of claim 12, wherein the neural network data of the second neural network model is obtained by training the second neural network model with audio samples associated with a target environment.

14. The computer system of claim 11, wherein the at least one computer processor is further configured to recognize and translate the clean speech signal into text.

1/8



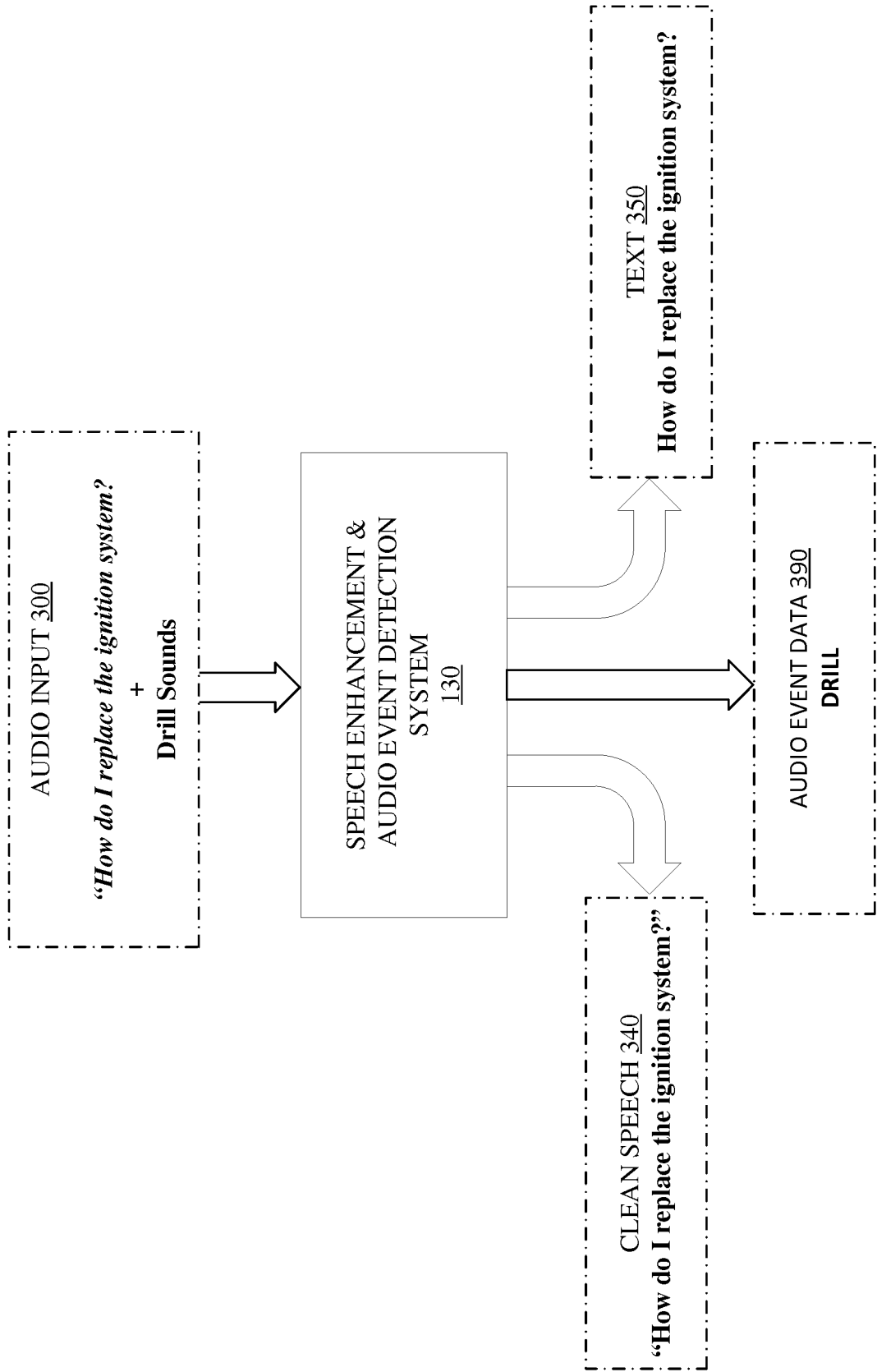


FIG. 2

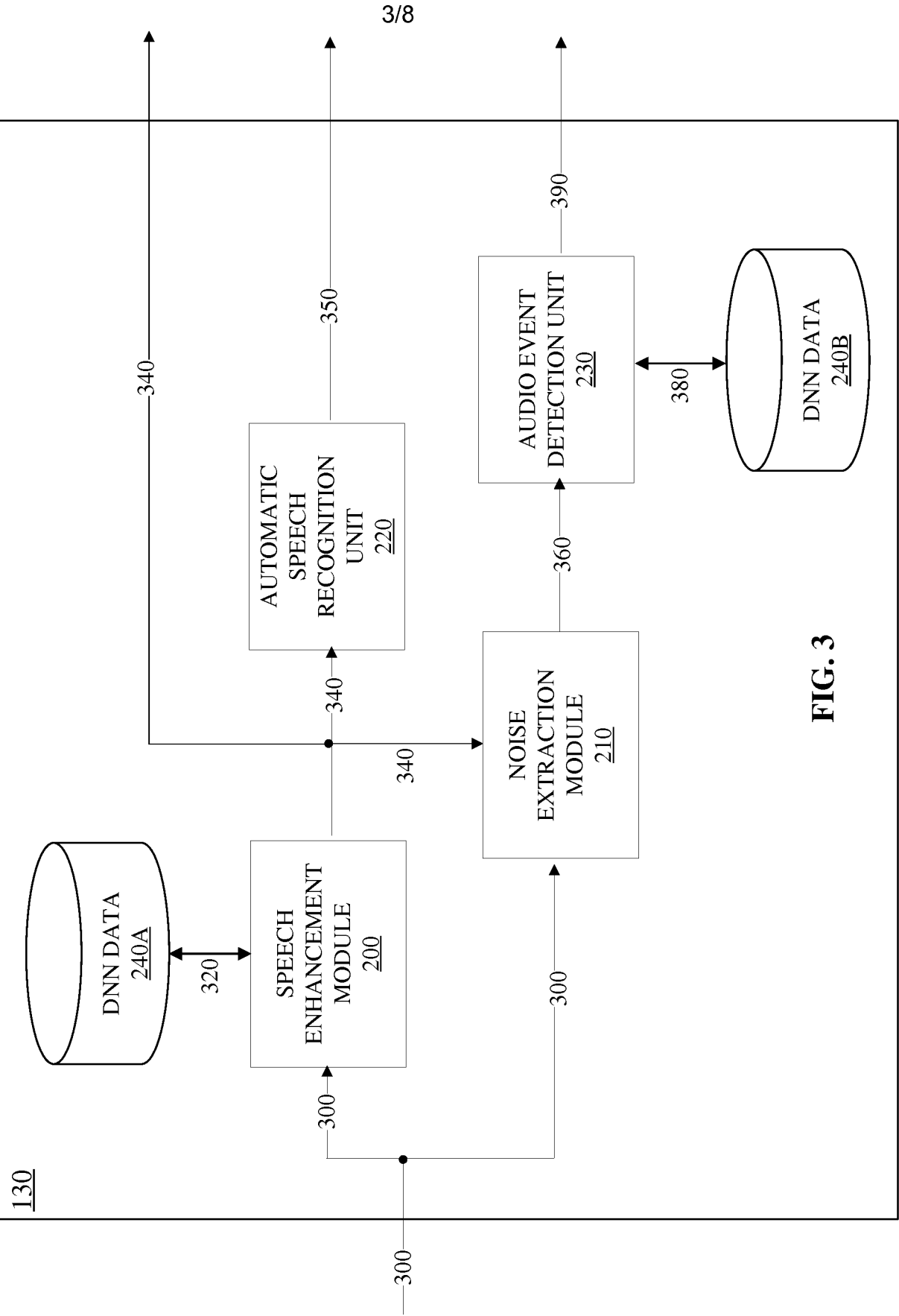


FIG. 3

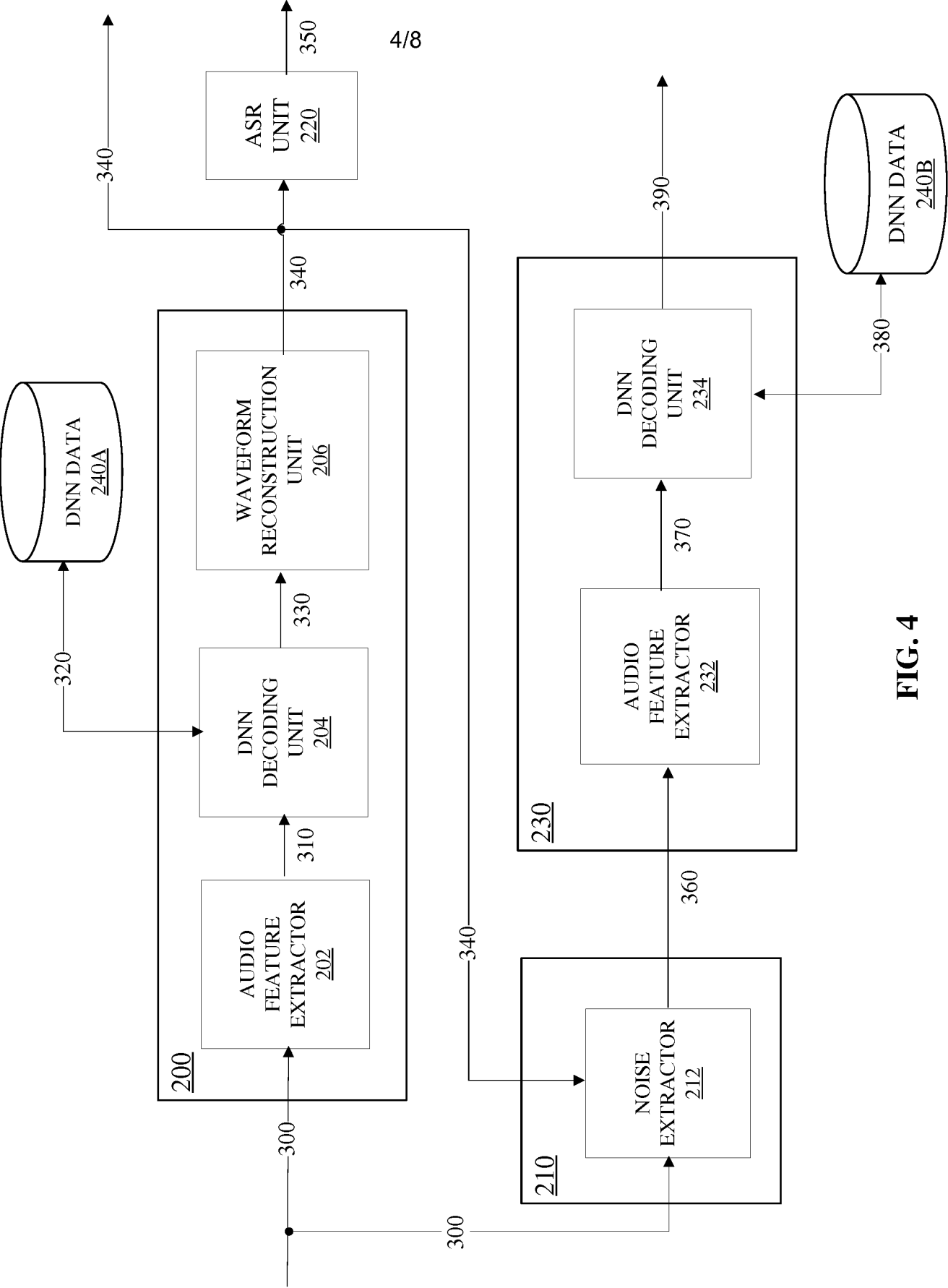


FIG. 4

5/8

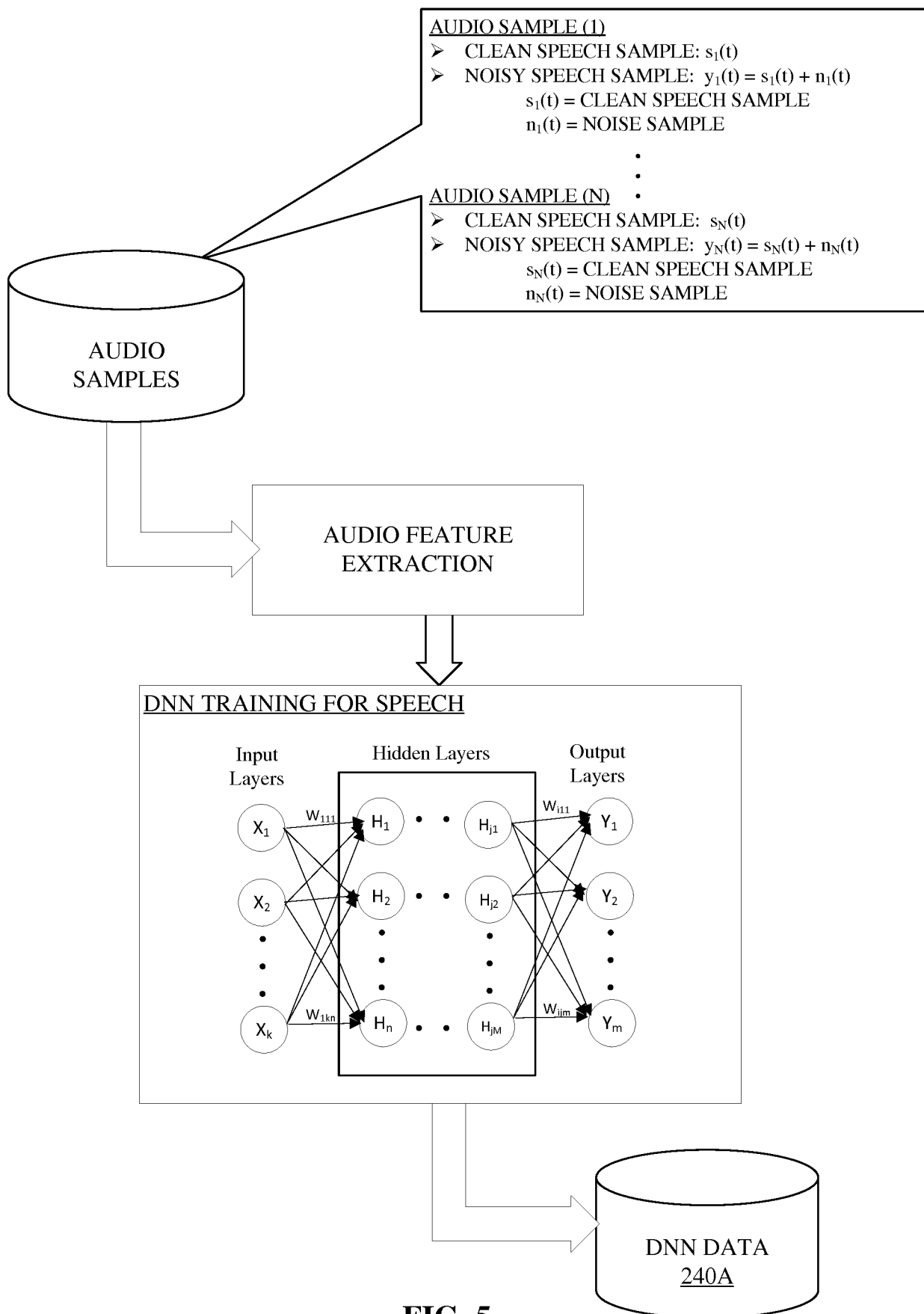
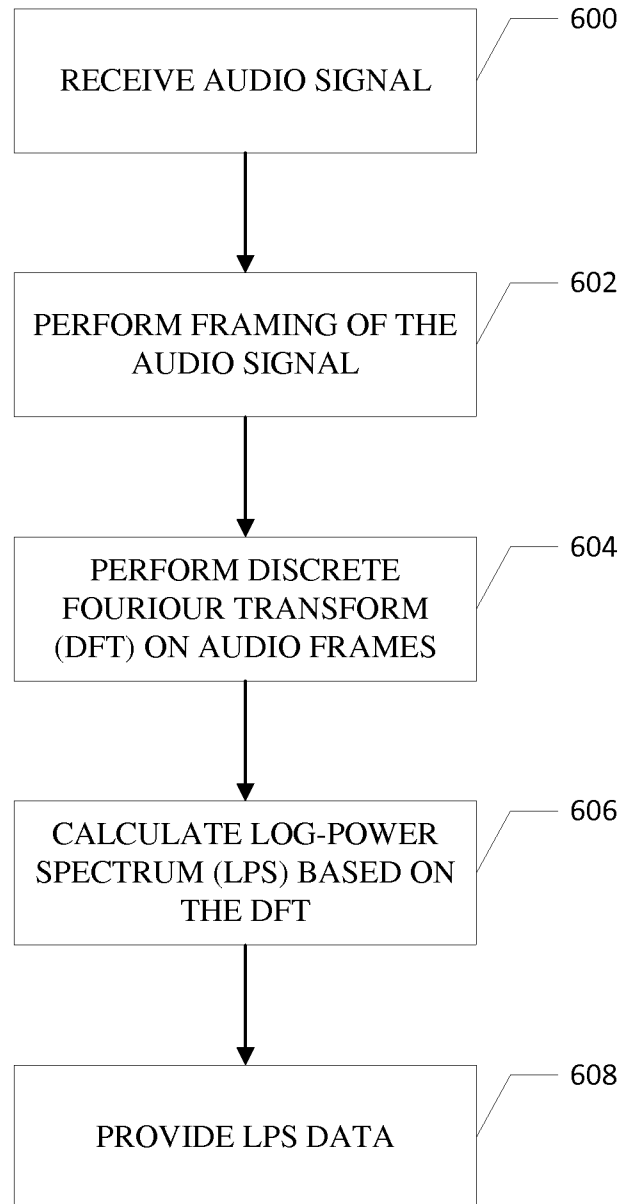


FIG. 5

6/8

**FIG. 6**

7/8

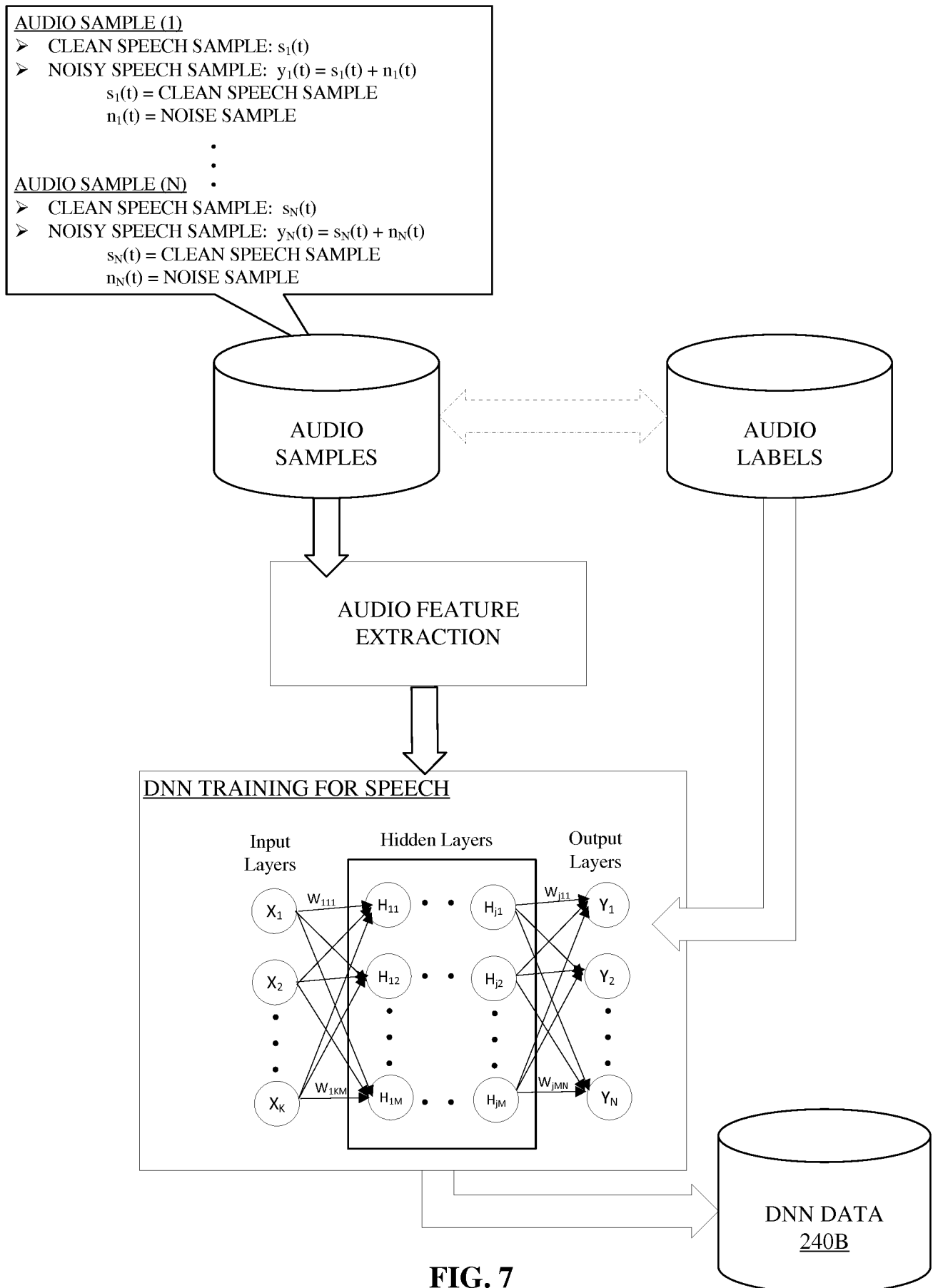
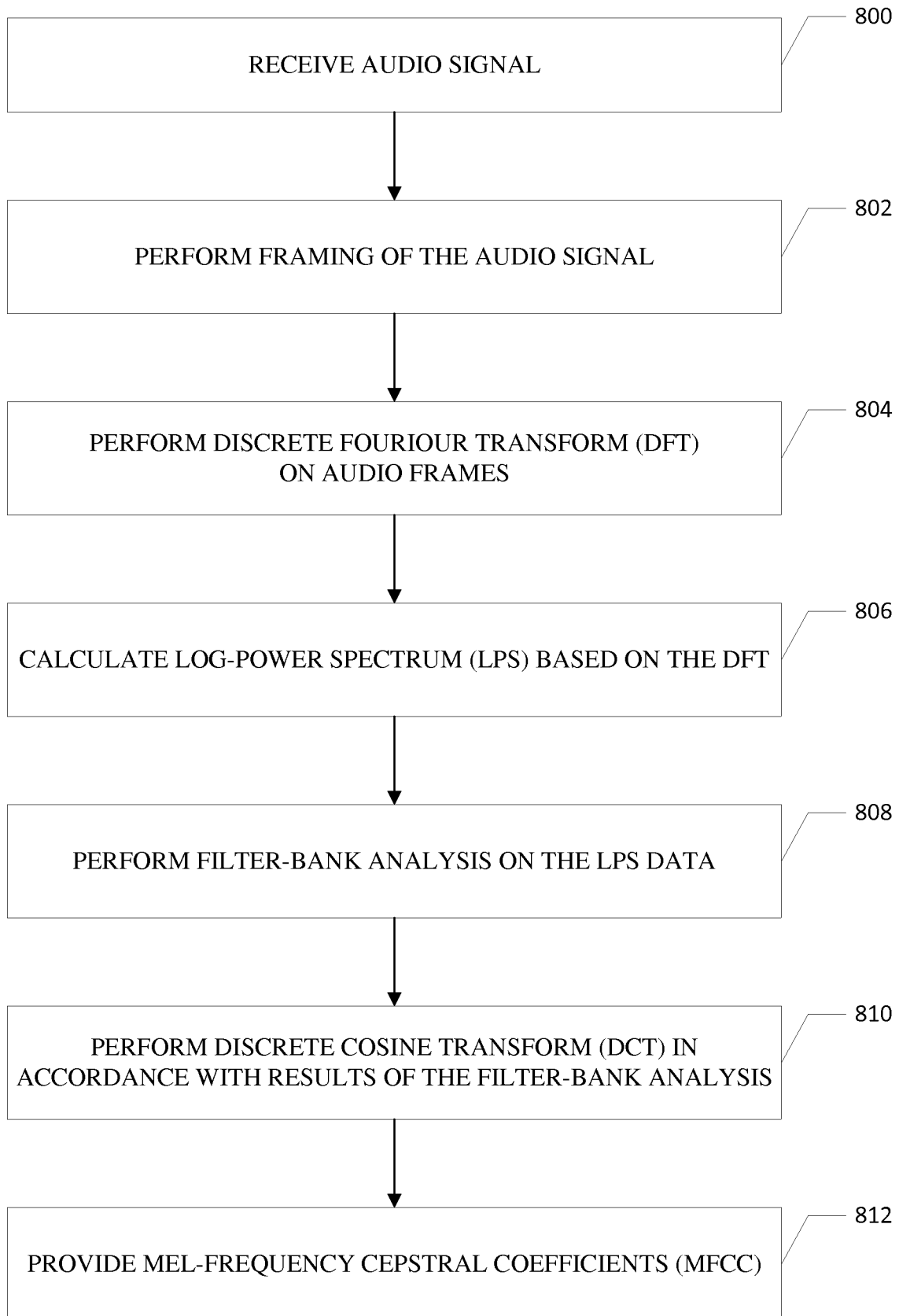


FIG. 7

8/8

**FIG. 8**

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/060637

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L21/0216

ADD. G10L25/30 G10L15/26

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2016/005422 A1 (ZAD ISSA SYAVOSH [US] ET AL) 7 January 2016 (2016-01-07)	1-3,8-10
A	figures 2,3,5 paragraphs [0011], [0013], [0016], [0017] paragraphs [0031], [0050], [0051] - [0081] paragraph [0048] paragraphs [0082], [0083], [0086] paragraph [0091] -----	4-7, 11-14
X,P	EP 3 057 097 A1 (NXP BV [NL]) 17 August 2016 (2016-08-17)	1-3,8-10
A,P	paragraphs [0004], [0010], [0011] paragraphs [0035], [0036], [0037], [0038], [0039] ----- -/-	4-7, 11-14



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

13 June 2017

Date of mailing of the international search report

22/06/2017

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Krembel, Luc

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/060637

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ESPI MIQUEL ET AL: "Spectrogram patch based acoustic event detection and classification in speech overlapping conditions", 2014 4TH JOINT WORKSHOP ON HANDS-FREE SPEECH COMMUNICATION AND MICROPHONE ARRAYS (HSCMA), IEEE, 12 May 2014 (2014-05-12), pages 117-121, XP032610791, DOI: 10.1109/HSCMA.2014.6843263 [retrieved on 2014-06-24] abstract paragraph [0003] -----	1-3,8-10
A	US 8 949 120 B1 (EVERY MARK [US] ET AL) 3 February 2015 (2015-02-03) column 3, lines 18-32 -----	4,11
A	EP 1 580 730 A2 (HARMAN BECKER AUTOMOTIVE SYS [CA]) 28 September 2005 (2005-09-28) abstract; figure 1 -----	5,6,12, 13

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2017/060637

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2016005422 A1	07-01-2016	CN 106663446 A	10-05-2017
		EP 3164871 A1	10-05-2017
		US 2016005422 A1	07-01-2016
		WO 2016004139 A1	07-01-2016

EP 3057097 A1	17-08-2016	CN 106024002 A	12-10-2016
		EP 3057097 A1	17-08-2016
		US 2016232915 A1	11-08-2016

US 8949120 B1	03-02-2015	NONE	

EP 1580730 A2	28-09-2005	CA 2501989 A1	23-09-2005
		CN 1737906 A	22-02-2006
		EP 1580730 A2	28-09-2005
		JP 2005275410 A	06-10-2005
		KR 20060044629 A	16-05-2006
		US 2006031066 A1	09-02-2006
